数据管理计划构成规范及其可操作数据监护模型研究*

刘　峰[1,2,3]　张晓林[1]
[1](中国科学院文献情报中心　北京　100190)
[2](中国科学院计算机网络信息中心　北京　100190)
[3](中国科学院大学　北京　100049)

摘要: 【**目的**】提出一套科学数据管理计划的细化构成规范; 并依此从可操作角度构建数据监护模型。【**方法**】对国际上主要科研管理机构的数据管理计划规范进行调研和统计; 并结合当前科研数据管理的需求与特点进行补充。【**结果**】形成 8 大基本构成要素和 39 个子要素的数据管理计划细化构成规范, 并构建出一种以数据管理计划为核心驱动的数据监护模型。【**结论**】数据管理计划细化构成规范可以完整、准确规范和指导科研数据的管理活动, 在操作层面上也可以有效地控制和约束科研全生命周期的数据监护过程。

关键词: 数据管理计划　数据监护　科研生命周期
分类号: G250

1　前　言

正如英国数据监护中心(Digital Curation Centre, DCC)总结的那样, 对科学数据进行有效的管理具有很多好处[1]。如: 可以在需要使用数据时更好地发现和理解数据; 在项目成员流动情况下保持项目的连续性; 避免数据重复采集或处理等重复性工作; 提供更多科研协作的机会; 提高个人的科研影响力; 增强科学数据的可追溯、可重用和可验证的能力等。而科学数据管理计划正是帮助和保证科研人员实现这些优点的最有效手段。

科学数据管理计划是概要介绍科研项目进行中以及结束后科学数据将如何被有效处理的正式文档,它不是固定不变的, 通常在项目的生命周期里被不断充实和细化完善从而变得更加准确[2]。通过科学数据管理计划科研人员可以全面掌握科研数据产生、处理、共享及应用的全过程, 方便地跟踪科研过程进展、并进行有针对性的决策; 进而保证科研全过程数据的有效管理。

近年来, 以数据为中心、数据驱动科研的特征越来越突出, 为保证科学研究的完整性, 国际上的基金组织和科研机构越来越重视科研项目中的科学数据管理与共享政策。例如: 自 2011 年 1 月开始, 美国国家科学基金会(National Science Foundation, NSF)[3]项目申请要求必须提交"数据管理计划(Data Management Plan, DMP)"; 英国研究理事会(Research Councils UK, RCUK)[4]下属 7 个研究理事会和惠康基金(Wellcome Trust)[5]分别发布了各自的数据管理计划要求; 地球数据观测网络(Data ONE)[6]、英国数据监护中心(DCC)[1]、美国政治与社会研究校际联盟(ICPSR)[7]等机构也都从各自角度提出细化的数据管理计划规范要求。

通讯作者: 刘峰, ORCID: 0000-0002-5816-2067, E-mail: liufeng@cnic.cn。

*本文系国家"十二五"科技支撑计划项目"农村信息服务云存储与云计算技术研究及平台建设"(项目编号:2013BAD15B02)、中国科学院"十二五"信息化项目"科学数据资源整合与共享工程"(项目编号:XXH12504)和中国科学院计算机网络信息中心一三五规划重点培育方向专项"科研大数据资源管理与服务平台及其关键技术"(项目编号:NIC_PY_1405)的研究成果之一。

2 构成要素分析

当前各机构发布的数据管理计划规范由于需求各异，主要构成要素之间有相似性又有差异性。这为完整而准确地理解科学数据管理计划提出了挑战。因此有必要全面梳理一下科学数据管理计划的构成。

为此笔者对上述主要的科研管理机构的数据管理计划进行细化统计，如表 1 所示：

表 1　主要科研管理机构数据管理计划基本要素统计

机构名称 / 要素名称	NSF[3]	Data ONE[4]	DCC[1]	ICPSR[7]	BBSRC[8]	Cancer Research UK[9]	MRC[10]	AHRC[11]	ESRC[12]	STFC[13-14]	Wellcome Trust[5]
	美国国家科学基金会	地球数据观测网络	英国数据监护中心	美国政治与社会研究校际联盟	英国生物技术与生物科学研究委员会	英国癌症研究中心	英国医学研究委员会	英国艺术与人文研究委员会	英国经济与社会研究委员会	英国科学技术设备委员会	惠康基金会
数据产生上下文	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
数据组织规范与策略	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
数据存储与安全管理	N	Y	Y	Y	N	N	Y	Y	Y	Y	N
数据道德规范与知识产权	N	N	Y	Y	N	N	N	N	Y	N	N
数据共享与服务实践	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
数据重用管理	Y	N	N	N	N	N	N	N	N	N	N
数据归档与长期保存	Y	N	Y	Y	N	Y	N	N	N	N	Y
数据资源保障计划	N	Y	Y	Y	N	N	N	N	N	Y	Y

(注: Y 表示包含此要素, N 表示不包含此要素)

经过对数据管理中数据的生成、组织、存储、权益保护、共享与重用、归档保存以及资源保障等方面的梳理与统计，最终归纳数据管理计划包括：数据产生上下文、数据组织规范与策略、数据存储与安全管理、数据道德规范与知识产权、数据共享与服务实践、数据重用管理、数据归档与长期保存和数据资源保障计划 8 个大类，其中：

①数据产生上下文：重点包括数据采集或生产的相关上下文(工具、环境、实验方法等)，以及未来用户可能的数据需求。

②数据组织规范与策略：确定如何将数据组织成文件，包括数据、元数据存储格式及标准，将会产生何种类型的数据产品。考虑何种关系数据库或其他数据组织策略更适合。

③数据存储与安全管理：明确数据存储地点、方式、管理责任链；备份、版本控制方式及工具；相关安全性、保密性等。

④数据道德规范与知识产权：重点包括数据资源所有数据版权和权限管理信息；明确提出数据共享的权益需求，包括数据使用许可、保密性及其他道德方面的考虑。

⑤数据共享与服务实践：描述共享数据的计划及数据访问发布方针，确定数据被发布条件、发布的范围、数据时滞期与目标日期、数据的发布共享方式、可用的分析工具以及如何引用等。

⑥数据重用管理：描述数据通常被访问或共享使用的策略。例如：免责声明的具体使用条款，数据在其他应用或产品中的应用条件等。

⑦数据归档与长期保存：标识出具有长期价值的数据集；描述说明用以保证关键数据集的长期价值的归档保存计划。

⑧数据资源保障计划：实现个人数据管理计划需要的资源信息；描述需要附加的经费、软件、硬件、是否需要技术专家支持与培训，以及如何能获取到这些资源。

3 细化要素分析

数据管理计划 8 个基本构成要素只是泛化说明了数据管理计划的基本结构，但要全面细致理解数据管理计划必须对构成要素进行细化分解，为此依据基本要素的内容，逐项对基本要素进行细化，形成一系列细化要素。同时结合当前科研数据管理的需求与特点，重点融入数据的保留期、时滞期、首发权声明及数据引用等子要素。并根据数据管理计划中要素的必备等级设置必选项、推荐项。整理的数据管理计划细化要素如表 2 所示。

表 2 数据管理计划细化要素分解

基本要素	细化要素	必备性	描述
数据产生上下文	数据来源	必选项	数据的来源信息描述，包括项目、科研背景信息等
	数据规模	必选项	数据存储量统计信息描述
	数据生产环境	必选项	数据采集或生产的环境上下文信息，如:地理位置、温度、湿度等
	数据生产工具	必选项	数据采集或生产的主要设备、仪器等工具描述
	数据生产方法	必选项	数据采集或生产的主要方法、过程描述
	质量控制与保证	推荐项	数据采集或生产的质量控制规范及方法描述
	潜在用户与需求	推荐项	数据未来潜在的用户及需求分析
数据组织规范与策略	数据类型	必选项	数据的主要存储类型描述
	存储格式	必选项	数据存储的主要标准格式名称信息
	元数据规范	必选项	元数据组织参照的标准、格式及组织形式
	数据组织策略	必选项	数据组织设计上的思路与策略
数据存储与安全管理	保留期	推荐项	数据的存储保留期限设置
	存储计划	必选项	数据的短期保存计划
	数据监护责任链	必选项	数据全生命周期管理各阶段细化责任人设置描述
	备份策略与工具	必选项	数据的短期备份策略、工具及方法
	版本控制方式与工具	推荐项	数据的版本控制方法、工具描述
	安全与保密规范	推荐项	数据存储上的安全与保密规范信息
数据道德规范与知识产权	伦理与隐私保护策略	必选项	数据在伦理与隐私保护方面的处理方法详细描述
	知情同意规范	推荐项	必要知情同意规范信息说明
	数据知识产权声明	必选项	数据的知识产权信息要求声明信息
	首发权声明	推荐项	数据的首发权要求声明信息
数据共享与服务实践	数据访问说明	必选项	数据访问的流程及方法信息
	数据访问接口	推荐项	数据访问的 API 开放接口信息
	时滞期	推荐项	数据共享发布前的时滞期设置信息
	数据共享计划	必选项	数据的开放共享发布计划
	共享说明	必选项	数据共享的范围、等级、条件和方式
	数据引用规范	必选项	数据引用规范与方法描述
	数据应用与服务	推荐项	数据的其他相关应用与服务流程与方式
数据重用管理	免责声明	推荐项	数据重用或二次发布的免责申明信息
	重用描述	必选项	数据重用或二次发布的条件及方式
	重用许可策略	必选项	数据重用或二次发布的许可策略描述
数据归档与长期保存	归档数据筛选	推荐项	需长期保存的重要数据的选择与描述
	归档保存计划	必选项	数据长期保存的归档保存计划
	长期保存规范	必选项	数据长期保存相关的组织、管理、引用、质量控制等规范
数据资源保障计划	经费预算计划	必选项	保障数据管理计划得以实现的经费预算计划
	软硬件环境信息	必选项	保障数据管理计划得以实现的软硬件环境需求
	培训与技术支持需求	推荐项	保障数据管理计划得以实现的人员培训与技术支持需求
	人员需求	必选项	保障数据管理计划得以实现的人员方面需求
	法律需求	推荐项	保障数据管理计划得以实现的法律方面需求

4 数据监护模型设计

从表 2 的细化要素分解中可以看出，数据管理计划覆盖了从数据产生、共享服务到数据归档保存的全生命周期监护的各个环节。它对科研中的数据管理提出了极为重要的规范和指导。从当前应用角度看，数据管理计划仍然停留在规范记录层面，为了进一步体现其指导、控制作用，笔者从可操作角度出发构建了以数据管理计划为核心驱动的科研全生命周期数据监护框架模型，如图 1 所示：

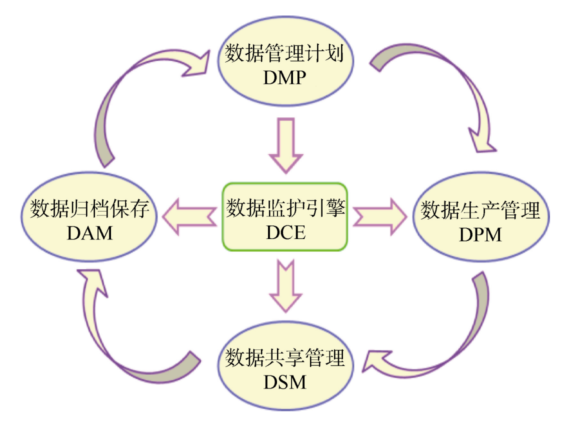


图 1 数据管理计划驱动的数据监护框架模型

(1) 数据监护引擎(Data Curation Engine, DCE)是科研数据全生命周期的核心控制模块，它负责约束和控制科研数据从数据生产、数据共享到数据归档保存的全生命周期管理的全过程。它又可细分为用户及权限管理、数据管理计划管理、工作流管理等子模块。

(2) 数据管理计划(Data Management Plan, DMP)构成了数据监护引擎的核心输入，其中数据管理计划的细化规范是构成数据监护引擎的核心要素。

(3) 数据生产管理(Data Production Management, DPM)、数据共享管理(Data Sharing Management, DSM)和数据归档保存(Data Archiving Management, DAM)是抽象出来的科研数据生命周期流转的重要功能运行模块，它们接受数据监护引擎的驱动和控制。每个模块又可划分为若干子模块。其中 DPM 包括：数据采集加工、数据组织、数据存储等子模块；DSM 包括：数据的权益控制、数据共享与发布等子模块；DAM 包括：数据归档备份、数据长期保存等子模块。

从整个框架模型可以看出数据管理计划是模型的核心驱动模块，其中数据管理计划规范是构成整个模型的核心要素。然而要真正实现数据管理计划的核心约束和控制作用，必须对细化规范中可量化控制要素进行过滤。表 3 展示了过滤后的数据管理计划可量化控制要素。

表 3 数据管理计划可量化控制要素过滤

基本要素	量化控制要素	要素说明
数据产生上下文	数据来源	可作为全局参数设置，供数据采集加工模块继承。
	数据生产环境	可作为全局参数设置，供数据采集加工模块继承。
	数据生产工具	可作为全局参数设置，供数据采集加工模块继承。
	数据生产方法	可作为全局参数设置，供数据采集加工模块继承。
数据组织规范与策略	数据类型	可作为全局参数设置，供数据组织模块继承。
	存储格式	可作为全局参数设置，供数据组织模块继承。
	元数据规范	可作为全局参数设置，供数据组织模块继承。
	数据组织策略	可作为全局参数设置，数据组织模块依此形式实现数据组织。
数据存储与安全管理	保留期	可作为全局参数设置，供数据存储模块控制数据的存储时限或归档时机。
	数据监护责任链	可作为全局参数设置，供数据存储模块进行权限控制。
	备份策略与工具	可作为全局参数设置，控制数据存储模块的备份执行方式。
	版本控制方式与工具	可作为全局参数设置，控制数据存储模块的版本控制执行方式。
数据道德规范与知识产权	伦理与隐私保护策略	可作为全局参数设置，供数据权益控制模块继承。
	知情同意规范	可作为全局参数设置，供数据权益控制模块继承。
	数据知识产权声明	可作为全局参数设置，供数据权益控制模块继承。
	首发权声明	可作为全局参数设置，供数据权益控制模块继承。

(续表)

基本要素	量化控制要素	要素说明
数据共享与服务实践	时滞期	可作为全局参数设置，供数据共享或发布模块进行数据发布时限控制。
	数据访问说明	可作为全局参数设置，供数据共享或发布模块继承并应用。
	数据访问接口	可作为全局参数设置，供数据共享或发布模块继承并应用。
	数据共享计划	可作为全局参数设置，供数据共享或发布模块继承并应用。
	共享说明	可作为全局参数设置，供数据共享或发布模块继承并应用。
	数据引用规范	可作为全局参数设置，供数据共享或发布模块继承并应用。
	数据应用与服务	可作为全局参数设置，供数据共享或发布模块继承并应用。
数据重用管理	免责声明	可作为全局参数设置，供数据共享或发布模块继承并应用。
	重用描述	可作为全局参数设置，供数据共享或发布模块继承并应用。
	重用许可策略	可作为全局参数设置，供数据共享或发布模块继承并应用。
数据归档与长期保存	归档数据筛选策略	可作为全局参数设置，供数据归档保存模块继承并应用。
	归档保存计划	可作为全局参数设置，供数据归档保存模块继承并应用。
数据资源保障计划	人员需求	可作为全局参数设置，供权限控制模块进行用户角色及权限的控制。

5 结　语

科学数据管理计划是有效规范数据管理全过程的文档，目前各个主要的数据监管机构从自身需求出发提出规范构成要求，但还没有形成统一的标准。这为科研项目中科学数据统一规范化管理提出了挑战。本文在参考主要数据管理机构数据管理计划规范的基础上，基于统计数据提出一个细化的科学数据管理计划构成规范，力图全面细化概括通用科学数据管理计划的主要要素及规范要求。

基于数据管理计划的细化规范的设计，笔者认为数据管理计划并不应该仅仅是简单的文本型记录规范，它的作用应更深刻地体现在其对科研数据管理的标准化和规范化指导与控制上。为此，本文从可操作角度提出以数据管理计划为核心驱动的科学数据监护模型。该模型通过对数据管理计划具体要素的量化设置，即可构建出一个细化控制驱动的科研全生命周期数据管理系统。

在国内，数据管理计划目前正逐渐被主要项目及数据管理机构所重视，同时国内外基于科研全生命周期的数据监护的研究正不断深入，希望本文的研究能为未来的科学数据管理实践提供有意义的参考与启发。

参考文献：

[1] Jones S. How to Develop a Data Management and Sharing Plan [EB/OL]. [2015-08-18]. http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How%20to%20Develop.pdf.

[2] Guidelines on Data Management in Horizon 2020 [EB/OL]. [2015-08-18]. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

[3] NSF: Proposal Preparation Instructions of DMP [EB/OL]. [2015-08-18]. http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp.

[4] Research Councils UK [EB/OL]. [2015-08-18]. http://www.rcuk.ac.uk/.

[5] Wellcome Trust: Guidance for Researchers: Developing a Data Management and Sharing Plan [EB/OL]. [2015-08-18]. http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers/index.htm.

[6] Strasser C, Cook R, Michener W, et al. Primer on Data Management: What You Always Wanted to Know [EB/OL]. [2015-08-18]. https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf.

[7] ICPSR: Elements of a Data Management Plan [EB/OL]. [2015-08-18]. http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/elements.html.

[8] BBSRC: Data Sharing Policy [EB/OL]. [2015-08-18]. http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf.

[9] Cancer Research UK: Data Sharing Guidelines [EB/OL]. [2013-08-18]. http://www.cancerresearchuk.org/funding-for-researchers/applying-for-funding/policies-that-affect-your-grant/submission-of-a-data-sharing-and-preservation-strategy/data-sharing-guidelines.

[10] MRC: Guidance on Data Management Plans [EB/OL]. [2015-08-18]. http://www.mrc.ac.uk/research/research-policy-ethics/

data-sharing/data-management-plans/.
[11] AHRC: Technical Plan [EB/OL]. [2015-08-18]. http://www.ahrc.ac.uk/funding/research/researchfundingguide/applicationguidance/technicalplan/.
[12] ESRC: Research Data Policy [EB/OL]. [2015-08-18]. http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf.
[13] STFC: Guidelines on DMPs [EB/OL]. [2015-08-18]. http://www.stfc.ac.uk/funding/research-grants/data-management-plan/.
[14] STFC: Scientific Data Policy [EB/OL]. [2015-08-18]. http://www.stfc.ac.uk/Resources/pdf/STFC_Scientific_Data_Policy.pdf.

作者贡献声明：

张晓林：提出围绕项目生命周期的监护引擎的概念，参与内容分析和组织讨论，审定论文；
刘峰：模型设计，收集、分析数据，论文撰写。

收稿日期: 2015-09-07

Research on the Specification of Data Management Plan and Its Operational Model

Liu Feng[1,2,3]   Zhang Xiaolin[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[3](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] Propose a set of detailed structure specifications of scientific data management plan and in accordance with a data curation model constructed from the operational perspective. [**Methods**] This paper carries on the research and the statistics on the scientific data management plan specification of the main research and management agencies in the world, and makes supplement combining with the requirement and characteristic of current scientific research data management. [**Results**] This paper forms the detailed structure specification of data management plan with 8 major basic elements and 39 sub-elements and constructs a data curation model taking data management plan as the core driver. [**Conclusions**] The detailed structure specification of data management plan may regulate and guide the activities of scientific data management completely and accurately, it can also be effectively controlled and restricted the data curation process of the whole life cycle of scientific research at the operational level.

Keywords: Data management plan   Data curation   Research lifecycle